VisToT: Vision-Augmented Table-to-Text Generation

Prajwal Gatti[1], **Anand Mishra**[1], **Manish Gupta**[2], **Mithun Das Gupta**[2]
[1]Indian Institute of Technology Jodhpur [2]Microsoft
{pgatti,mishra}@iitj.ac.in, {gmanish,migupta}@microsoft.com

Abstract

Table-to-text generation has been widely studied in the Natural Language Processing community in the recent years. We give a new perspective to this problem by incorporating signals from both tables as well as associated images to generate relevant text. While tables contain a structured list of facts, images are a rich source of unstructured visual information. For example, in the tourism domain, images can be used to infer knowledge such as the type of landmark (e.g., church), its architecture (e.g., Ancient Roman), and composition (e.g., white marble). Therefore, in this paper, we introduce the novel task of Vision-augmented Table-To-Text Generation (VisToT), defined as follows: given a table and an associated image, produce a descriptive sentence conditioned on the multimodal input. For the task, we present a novel multimodal table-to-text dataset, WikiLandmarks, covering 73,084 unique world landmarks. Further, we also present a competitive architecture, namely, VT3 that generates accurate sentences conditioned on the image and table pairs. Through extensive analyses and experiments, we show that visual cues from images are helpful in (i) inferring missing information from incomplete or sparse tables, and (ii) strengthening the importance of useful information from noisy tables for natural language generation. We make the code and data publicly available[1].

1 Introduction

Structured data-to-text generation is a well-studied problem that demands a model to produce meaningful and factually accurate sentences based on its comprehension of the source context that are usually in the form of fact graphs, tables, or hierarchical spreadsheets (Lebret et al., 2016; Wiseman et al., 2017). Many datasets and several competitive models have been proposed for data-to-text generation in the last few years. Data-to-text has been used to generate variety of texts including weather reports (Angeli et al., 2010), sports news (Wiseman et al., 2017) and biographies (Lebret et al., 2016).

Conventional Transformer-based Table-to-Text:
"Lough Leane is a 4700 acre estate in Killarney, County Kerry, Ireland."

Vision-augmented Table-to-Text (this work):
"Lough Leane is a large lake in Killarney, County Kerry, Ireland."

Ground Truth: "Lough Leane is the largest of the three lakes of Killarney, in County Kerry."

Figure 1: VisToT: Vision-augmented Table-to-Text. In this novel proposed task, the goal is to explore the utility of visual cues besides the information in the table for natural language generation. Note the presence of large lake in the text generated by our proposed model; on the other hand, existing table-to-text models wrongly generate estate in the output text.

Specifically, for table-to-text, several benchmark datasets have been proposed including WikiBio (Lebret et al., 2016), WebNLG (Gardent et al., 2017), E2E (Novikova et al., 2017), DART (Nan et al., 2021), RotoWire (Wiseman et al., 2017), and Wikipedia Person and Animal dataset (Ye et al., 2020). These datasets vary in terms of their lexical richness, syntactic variation, and semantic and linguistic adequacy. Further, there are several recent methods utilizing neural encoder-decoder frameworks to generate text descriptions from tables (Lebret et al., 2016; Bao et al., 2018; Chisholm et al., 2017; Liu et al., 2018). Despite this progress, the current table-to-text literature has underexplored the challenge of including a multimodality requirement in this task.

While tables provide a series of facts at the conceptual level, images provide unstructured visual-level signals. Therefore, the information in the two modalities can be complementary and may lead

[1]https://vl2g.github.io/projects/vistot

Dataset	Domain	#Data-Text Pairs	Has Images?
WeatherGov (Liang et al., 2009)	Weather	22.1K	✗
Wikibio (Lebret et al., 2016)	Biography	728.3K	✗
RotoWire (Wiseman et al., 2017)	Baseketball	4.9K	✗
WebNLG (Gardent et al., 2017)	Selected DBPedia Categories	25.3K	✗
E2E (Novikova et al., 2017)	Restaurants	50.6K	✗
LogicNLG (Chen et al., 2020a)	Open Domain	37.0K	✗
ToTTo (Parikh et al., 2020)	Open Domain	136.2K	✗
WIKILANDMARKS (Ours)	**World Landmarks**	**73.2K**	✓

Table 1: Comparison of Data-to-Text generation datasets. WIKILANDMARKS is the only dataset with tables along with associated images.

to significant improvements in generated text, especially for domains such as tourism, dining, and products, where images contain rich information. Images could also act as a significant source of additional information when tables are sparse or noisy. Hence, in this paper, we propose an extension to the typical table-to-text task, where not just tables but also associated images are needed to generate meaningful and complete sentences. We refer to this novel task as Vision-Augmented Table-to-Text Generation or VISTOT. Figure 1 shows an example for this task. Note the presence of "large lake" in the text generated by our proposed model; on the contrary, existing table-to-text models wrongly generate "estate" in the output text.

Unlike table-to-text datasets that only contain (table, text) pairs, VISTOT requires a dataset that maps (table, image) pair to a text. Hence, we contribute a new dataset – WIKILANDMARKS, with infoboxes (tables) corresponding to 73,084 unique landmarks, ∼10 images on average per landmark, and a descriptive text with ∼35 tokens on average. Table 1 shows a comparison of various datasets with WIKILANDMARKS.

Further, to perform VISTOT on WIKILANDMARKS, we first experiment with standard previously proposed methods of encoding table and images. Next, to handle the joint tables and image information more accurately, we propose a novel model, namely Visual-Tabular Data-to-Text Transformer or VT3. For encoding image information, we experiment with Faster RCNN (Ren et al., 2015), ViT (Dosovitskiy et al., 2020), CLIP-ViT (Radford et al., 2021a) and Swin (Liu et al., 2021). For encoding tables, we represent key-value pairs in a sequence. Both the table and visual encodings are fed as input to our model initialized from the pretrained BART (Lewis et al., 2020) to generate text. To adapt the proposed model for our task, we present three novel pretraining strategies, namely image-table matching, masked value modeling, and image captioning.

In summary, our contributions are as follows: (i) A novel task namely VISTOT, of natural text generation from the table and image data. (ii) An accompanying multimodal dataset in English viz. WIKILANDMARKS having ∼73K samples with a focus on the domain of world landmarks. (iii) A novel method, VT3, that performs natural language text generation conditioned on multimodal data. (iv) Extensive automatic and human evaluations reporting results on our dataset.

2 Related Work

2.1 Table-to-Text Generation

Data-to-text generation is a long-established problem with the objective of generating sentences from structured data. Early works (Sripada et al., 2003; Reiter, 2007; Liang et al., 2009) focused on template-based data-to-text generation. WikiBio (Lebret et al., 2016), Rotowire (Wiseman et al., 2017), and E2E (Novikova et al., 2017) proposed more challenging and diverse datasets focusing on domains of biographies, basketball games, and restaurants, respectively. Modern works such as ToTTo (Parikh et al., 2020) and WebNLG (Gardent et al., 2017) go a step beyond and cover open-domain data, including many categories of world knowledge, as well as introduce data in graphical form. More recent methods deal with logical, numerical, and hierarchical reasoning in domains like sports, politics, entertainment (Chen et al., 2020a) and products (Zhang et al., 2022). While such works have become popular benchmarks for evaluating competitive table-to-text generation methods (Lewis et al., 2020; Su et al., 2021), current literature has underexplored the challenge of including a multimodality requirement in this task. In this work, we propose an extension to the current table-to-text task, where besides tables, associated images are also useful to generate meaningful and complete sentences. We believe that the data-to-text generation community would benefit from this challenge of vision augmentation to the rather classical and well-explored unimodal task.

2.2 Vision-and-Language Models

Besides vision and language tasks being explored independently, steady progress has been made toward problems that require joint modeling of both

Property	Value
Training set size	58,456
Unique landmarks/infoboxes	73,084
Images per landmark (Median/Avg)	4/10.0
Avg Target Length (tokens)	34.7
Target vocabulary size	160,203
Fields per infobox (Median/Avg)	12/13.0
Validation set size	7,314
Test set size	7,314

Table 2: WIKILANDMARKS dataset statistics.

modalities. These joint Vision-and-Language (V-L) tasks demonstrate visual understanding by generating or responding to language in the context of images or videos. Multimodal Transformers like VisualBERT (Li et al., 2019), ViLBERT (Lu et al., 2019), LXMERT (Tan and Bansal, 2019) and CLIP (Radford et al., 2021b) have shown impressive performance on several V-L tasks such as visual question answering, visual commonsense reasoning, and text-to-image retrieval by following a "pretrain-then-finetune" approach. We follow a similar approach to model table and image information jointly by devising novel task-specific pretraining strategies, namely image-table matching, masked value modeling, and image captioning.

3 The VISTOT Problem and Accompanying Dataset

3.1 VISTOT: Problem Definition

The proposed Vision-augmented Table-to-Text or VISTOT task can be formally defined as follows. Given a labeled training dataset $\mathcal{D} = \{\langle t_j, i_j, s_j \rangle\}_{j=1}^{N}$ where t_j is a table with key-value pairs describing an entity e, i_j is a related image and s_j is a natural language description for e containing m_j words $w_1, w_2, \ldots, w_{m_j}$, the goal is to learn a model with parameters θ^* such that the probability P of generating sentence s_j, given inputs t_j and i_j, is maximized. Mathematically,

$$\theta^* = \arg\max_{\theta} \prod_{j=1}^{N} \prod_{k=1}^{m_j} P(w_k | c_k, t_j, i_j, \theta)$$

Here $c_k = w_1, \ldots, w_{k-1}$ is the sequence of all context words preceding w_k.

In this work, we choose to work on the landmarks domain as descriptions of landmarks usually have important visual aspects such as architecture, type of building, composition, and surroundings. Nevertheless, our proposed model is equally suitable for all domains where visual inputs provide complementary information to factual information in the table for natural language generation.

3.2 WIKILANDMARKS: A novel dataset for Vision-augmented Table-to-Text

We introduce WIKILANDMARKS – a dataset for studying the novel problem of vision-augmented Table-to-Text in the English language. In the literature, there are several datasets for studying the table-to-text task, such as WikiBio (Lebret et al., 2016), Rotowire (Wiseman et al., 2017), ToTTo (Parikh et al., 2020). We could have used these datasets by augmenting corresponding images to them. However, the images are unlikely to have a greater impact in generating text corresponding to the tables in these datasets. For example, it might be hopelessly hard to improve the biographical summary of a person by adding their images to the Wikibio dataset. Therefore, we curate a new dataset where table-to-text can benefit from using relevant visual cues. To this end, we present WIKILANDMARKS, which contains 73,084 tables augmented with 766,723 images of world landmarks, and a brief text summary (first sentence of the Wikipedia webpage) corresponding to each table. We make code and data publicly available[2].

Data curation: We begin our data collection by obtaining a list of prominent world landmarks from the Google Landmarks Dataset v2 (Weyand et al., 2020). We then harvest Wikipedia pages and infoboxes corresponding to these landmarks. We remove all those landmarks which either do not contain infoboxes or contain completely non-English Wikipedia pages[3]. Each data sample of WIKILANDMARKS contains an infobox, a landmark image and a Wikipedia summary.

Dataset split: The final dataset has been divided into: training (80%), validation (10%) and test (10%) sets by ensuring non-overlap of entities or landmarks between the sets. A summary of dataset statistics is reported in Table 2.

Figure 3 shows the frequency distribution of the number of key-value pairs across tables in our dataset. The average number of key-value pairs is 13.0, with a standard deviation of 5.4. The top-15 most frequent keys in the dataset in non-ascending order are: name, location, image, coordinates, caption, website, architect, type, area, built, photo, country, map_caption, locmapin, map established.

[2] https://vl2g.github.io/projects/vistot

[3] Current version of WIKILANDMARKS only focuses on generating English text.

Figure 2: Word cloud for top few values for architectures, countries, materials, and types for the landmarks respectively (left to right) in the WIKILANDMARKS dataset.

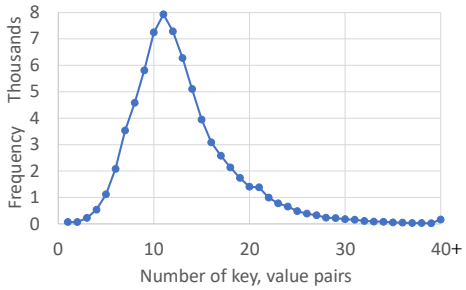


Figure 3: Frequency Distribution of Number of key-value pairs per table in WIKILANDMARKS

Category	Example	Percentage
Type of landmark	Castle, Statue	87
Type of holy place	Mosque, Pagoda	25
Architectural style	Neo-classical, Mughal	21
Type of building	Office, Museum	5
Composition of landmark	Steel, Marble	5
Type of park	Urban, Sports	5
Is the landmark a ruin?	Yes / No	4

Table 3: Distribution of types of visual information found among 100 randomly chosen target sentences from WIKILANDMARKS. Note that the percentage values do not add up to 100 because sentences could have overlap of categories.

Figure 2 shows word clouds of frequent values of popular visual attributes in the dataset.

Further, we scrutinized 100 randomly-chosen sentences from WIKILANDMARKS to understand the extent and type of visual information present. Table 3 shows the distribution of types of visual information. We observe that there is good hope of extracting visual cues from landmark images and using them for improving text generation. We provide further analyses of WIKILANDMARKS in the Appendix.

4 VT3: Visual-Tabular Data-to-Text Transformer

To address VISTOT, we present a novel trainable neural architecture namely Visual-Tabular Data-to-Text Transformer (or VT3 in short). VT3 is a BART-based encoder-decoder Transformer model. The overall architecture of VT3 and our proposed novel pretraining tasks are illustrated in Figure 4. Next, we describe the architectural details of VT3 followed by proposed pretraining strategies.

4.1 Image Encoding

An image embedding approach is required to feed images as input to our model. A common approach among several popular VL-models (Tan and Bansal, 2019; Li et al., 2020; Chen et al., 2020b) is to use region features of the images, also referred to as bottom-up features (Anderson et al., 2018). These capture semantic features of salient objects in images. They are obtained from detectors like Faster-RCNN (Ren et al., 2015), which is trained on the Visual Genome (Krishna et al., 2017) dataset to detect common objects (e.g., couch, dog) in images. However, region features have these limitations: (i) RCNN models are limited to express only a predefined set of object categories in images. (ii) Landmark images do not often contain common objects such as couches or dogs, and their presence is irrelevant to the landmarks.

Hence, we also experiment with a much simpler, lightweight, and convolution-free approach of using a Vision Transformer to embed images. Vision Transformers embed the whole image rather than a few select regions of interest. In particular, we experiment with ViT (Dosovitskiy et al., 2020), CLIP-ViT (Radford et al., 2021a) and Swin-Transformer (Liu et al., 2021) to extract a sequence of grid features from non-overlapping patches of the image. Given an image I, we first use these methods to obtain a set of g grid features, $\{x_i\}_{i=1}^{g}$. We set $g = 12 \times 12$ in our experiments and use a linear layer to transform each grid representation x_i to

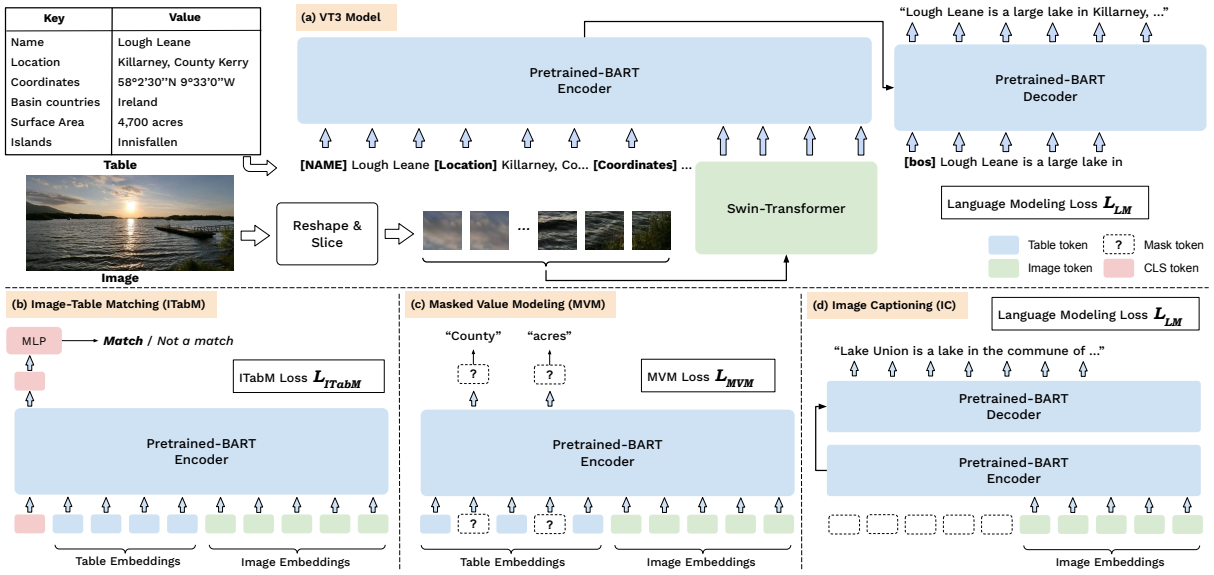


Figure 4: **Overview of the proposed Visual-Tabular Data-to-Text Transformer viz. VT3**: (a) Given a table (T) and an associated image (I) of a landmark entity, VT3 aims to generate an accurate summary text. Encoded visual features are obtained using Swin. However, other visual encoders such as Faster-RCNN, ViT, or CLIP-ViT can also be used. The multimodal transformer processes the encoded table and visual features to autoregressively generate a summary text. (b), (c) and (d) illustrate the pretraining tasks of VT3, namely Image-Table Matching (ITabM), Masked Value Modeling (MVM), and Image Captioning (IC), respectively. These pretraining objectives aid in learning richer multimodal alignment between landmark tables and images.

match the dimensions of the BART input. Overall, the image I is encoded as $F_I = \{linear(x_i)\}_{i=1}^{g}$.

4.2 Table Encoding

In the literature, several ways have been explored to effectively encode tabular information (Lebret et al., 2016; Liu et al., 2018). We note, however, that Wikipedia infoboxes are rather simple tables with a list of entries or fields in the form of key-value pairs. A straightforward approach to encoding such tables would be to list their key-value pairs as a sequence and feed it to an encoder. Inspired by Su et al. (2021), we represent each key as a special token in the model's vocabulary to better capture the semantic meaning of unique keys in the WIKILANDMARKS dataset. The embeddings for these keys are learned from scratch.

Thus, given a Wikipedia infobox table T with n key-value pairs $\{\langle k_1, v_1\rangle, \langle k_2, v_2\rangle, \ldots, \langle k_n, v_n\rangle\}$, we encode its keys $\{k_i\}_{i=1}^{n}$ as special tokens $\{\kappa_i\}_{i=1}^{n}$, and the values are tokenized using the byte-level Byte-Pair-Encoding scheme of the BART-tokenizer. Each of the key-value pairs are separated using an end-of-field (EOF) token. Thus, the table representation input to BART is $F_T = [\kappa_1, v_1, \text{EOF}, \kappa_2, v_2, \text{EOF}, \ldots, \kappa_n, v_n]$.

We feed the image representation F_I along with the table representation F_T to a BART encoder which ensures joint processing of semantics across the image and the table. Finally, we unify the visual (F_I) and tabular (F_T) input by concatenating them sequentially, with a [SEP] token between them, as $E = [F_I, \text{SEP}, F_T]$. Furthermore, positional embeddings are added to preserve sequential information, and different segment embeddings are added to the table and visual embeddings to further aid the model in differentiating the two modalities.

4.3 Pretraining Strategies

As VT3 is initialized from the pretrained BART, a language model, it lacks joint modeling for vision and language data. To strengthen the relationship between these two modalities and learn richer features to aid the VISTOT task, we propose three novel pretraining objectives to be used in a multi-stage manner, as described next.

4.3.1 ITabM: Image-Table Matching

Analogous to the Image-Text Modeling objective popular in vision-language pretraining, we propose Image-Table Matching (ITabM). In ITabM, the VT3 encoder is presented an image-table pair as [CLS F_T SEP F_I] and tasked to predict whether the table and image are a matched pair, i.e., whether they describe the same landmark. The output of

Method	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BLEURT
Image captioning-based						
PureT (Wang et al., 2022)	6.4	26.1	33.2	12.8	31.1	0.40
Table-to-Text						
Pointer-Generator (See et al., 2017)	17.8	39.2	51.6	31.7	49.2	0.50
BERT-to-BERT (Rothe et al., 2020)	22.1	43.9	55.3	35.6	53.1	0.50
T5 (Kale and Rastogi, 2020)	25.8	48.1	58.8	38.8	57.0	0.54
PlanGen (Su et al., 2021)	8.6	20.6	32.5	20.2	31.9	0.49
Visual-Tabular Data-to-text						
LSTM+ResNet50	6.5	19.8	31.0	19.1	30.3	0.39
VisualBERT+BERT	26.1	49.0	60.4	39.2	58.8	0.54
VT3	**30.2**	**53.5**	**62.9**	**43.4**	**60.8**	**0.56**

Table 4: Performance comparison across various methods on WIKILANDMARKS test set. Details in Section 5.

the classification token, h_{CLS}, is taken to indicate the fused representation of both modalities, and an MLP layer is learned to predict the match score s. We randomly replace the table or image with another to generate negative pairs. We utilize the binary cross-entropy loss to train ITabM.

4.3.2 MVM: Masked Value Modeling

In this objective, we mask ∼15% of the values part of key-values in the infobox tables and replace them with a special token [MASK]. We then train the model to reconstruct the masked tokens given the context of the image and remaining table input. The Masked Value Modeling objective is inspired by the Masked Language Modeling objective in BERT (Devlin et al., 2018). The model predicts a likelihood distribution over the vocabulary for the masked token and is trained to minimize the negative log-likelihood loss. We mask table values as they are more contextually relevant than keys. Values also contain nouns such as "church" and "cottage", that are inferable from the image, unlike table keys such as "Location" and "Name".

4.3.3 IC: Image Captioning

The Image Captioning objective trains the model to maximize the likelihood of a target sentence using only the image of a landmark. It follows the same language modeling loss that we finetune the VT3 model with. This pretraining objective helps the model learn to interpret visual features of various landmark categories and their attributes and accurately describe them in the generated text.

VT3 is pretrained on all the three training objectives on WIKILANDMARKS itself. We empirically found that the multistage pretraining approach works best for the objectives. First, the model is trained with ITabM, followed by MVM and IC, respectively, and then fine-tuned for the VISTOT task.

Metric	FRCNN	CLIP-ViT	ViT	Swin
BLEU	27.4	28.2	29.6	**30.2**
METEOR	50.8	51.6	52.9	**53.5**
ROUGE-1	59.9	60.3	61.7	**62.9**
ROUGE-2	42.3	43.0	42.7	**43.4**
ROUGE-L	58.2	58.9	59.5	**60.8**

Table 5: Ablation with different VT3 visual encoders.

5 Experiments

5.1 Baselines and Metrics

We compare with the following baseline methods.

Image Captioning-based approach. It may not be trivial to generate sentences in WIKILANDMARKS using images alone. However, to have a comprehensive evaluation, and assess the importance of visual signals in our task, we compare our approach against a state-of-the-art image captioning-based baseline, PureT (Wang et al., 2022).

Table-to-Text Approaches. Table-to-text is a well-explored area in the literature. We compare our approach against the following four recent approaches: Pointer-Generator (See et al., 2017), BERT-to-BERT (Rothe et al., 2020), T5 (Kale and Rastogi, 2020) and PlanGen (Su et al., 2021). These baselines help us understand what performance can be achieved on WIKILANDMARKS by using only the table and no associated images.

Visual-Tabular Data-to-Text Approaches. We also propose and compare with baseline models that use images as well as tables to generate text similar to our proposed work. Specifically, we compare against: (i) LSTM+ResNet50 model: LSTM is used to encode the table, and ResNet50 to encode the image, followed by late fusion. The decoder is also an LSTM. (ii) VisualBERT+BERT Model: Transformer method with VisualBERT-initialized encoder paired with a BERT-initialized decoder.

We use the standard natural language generation and image captioning metrics such as BLEU, ME-

Model	Image	Table	Pretraining	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
1	×	✓	×	25.3	47.0	58.4	38.9	56.8
2	✓	×	×	3.9	19.9	26.8	8.1	24.7
3	✓	✓	×	27.7	50.0	60.8	41.2	59.1
4	✓	✓	MVM	28.1	50.9	61.4	42.0	60.2
5	✓	✓	ITabM	28.4	51.4	61.7	41.9	59.6
6	✓	✓	IC	29.5	51.9	62.0	42.4	60.1
7	✓	✓	ITabM + MVM + IC	**30.2**	**53.5**	**62.9**	**43.4**	**60.8**

Table 6: Ablation studies for the VT3 model on the WIKILANDMARKS test set. Model 1 is the BART model finetuned only on the table part of WIKILANDMARKS. Thus, model 1 is VT3 (w/o vision).

	VT3 (w/o vision)					VT3				
Removed Key	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
T	25.0	46.8	58.2	38.6	56.6	29.4	52.8	62.1	42.6	60.0
A	25.2	47.0	58.4	38.9	56.8	29.7	53.0	62.4	42.9	60.3
C	25.3	47.0	58.4	38.9	56.8	30.0	54.4	62.7	43.3	60.6
T+A	25.0	46.7	58.1	38.6	56.6	29.0	52.3	61.7	42.1	59.6
T+C	25.0	46.8	58.2	38.6	56.6	29.2	52.7	62.0	42.5	59.9
A+C	25.2	47.0	58.4	38.9	56.8	29.6	53.0	62.3	42.8	60.2
T+A+C	24.9	46.7	58.2	38.6	56.6	28.9	52.2	61.6	42.0	59.5
Random	25.2	47.0	58.4	38.9	56.8	29.8	53.0	62.5	43.0	60.4

Table 7: **Missing Keys Experiment.** We probe VT3 for its usage of visual modality by masking few visually strong keys in the infobox. Here, T=Type, A=Architecture, C=Composition.

TEOR, ROUGE-1, ROUGE-2 and ROUGE-L, and BLEURT for evaluating the performance of text generation in our experiments. Higher values for all the scores are desired. Further, to measure how humans perceive the quality of the generated text, we also report human evaluation using fluency and faithfulness defined in Section 5.3. Implementation details for reproducibility are detailed in Appendix B.

5.2 Quantitative Results and Ablations

We compare the accuracy of our proposed model, VT3, against the baselines in Table 4. We observe that the image captioning-based approach performs poorly. This result is obvious as predicting precise facts such as location, date of inception, etc., are non-trivial and nearly impossible from images alone. Table-to-text baselines achieve better performance on VISTOT. However, as image-blind models, table-to-text models fall short. As expected, the visual-tabular data-to-text models perform the best. Overall, we observe that on all the automatic performance measures, VT3 significantly outperforms the most competitive table-only models and other visual-tabular models. Particularly, LSTM+ResNet50 performs poorly because of a lack of pretraining, inability to handle out-of-vocabulary tokens, and inability to capture interactions between table and images due to late fusion.

We perform ablations on VT3 to study the impact of (i) visual encoder and (ii) pretraining strategies. We investigate the impact of four strong visual encoders on VISTOT in Table 5. We observe that Faster-RCNN, an object-detection-based vision encoder performs the poorest, which is expected as visual objects, and their relationships do not play a strong role in our data. Further, recently introduced Vision Transformers (ViT, CLIP-ViT, and Swin) perform well, and Swin outperforms all encoders. We posit that its effectiveness is due to its hierarchical architecture, allowing it to capture features more accurately at different scales.

Further, in Table 6, we perform an ablation to illustrate the individual and combined efficacy of our proposed pretraining strategies (please refer to Section 4.3). IC approach is observed to be the most effective pretraining method. We theorize that since the model is trained to generate sentences based on the images alone in IC, it learns to strongly utilize the visual information to generate better captions in the VISTOT task.

Missing Keys Experiment: To further test the utility of visual signals in cases when certain keys are missing from the table, we perform an experiment by removing keys of Type, Architecture, Composition, their combination, and random keys. Under these ablation settings, we compare the results using the full VT3 model and the VT3 model without the image input, i.e., VT3 (w/o vision).

We show this result in Table 7. In all the combinations, VT3, by virtue of having access to the visual information, surpasses VT3 (w/o Vision).

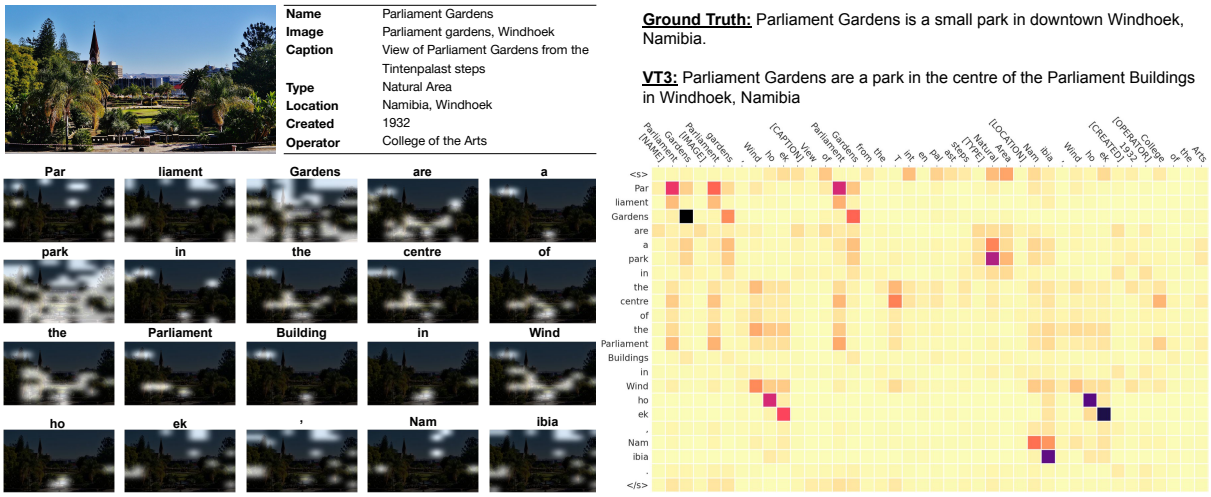


Figure 5: Multimodal Attention Visualization during sentence generation. We observe that to generate the words 'Garden' and 'park,' VT3 attends to the associated regions of the image as well as the relevant key 'Type' in the table. Further, we see that the model attends to the building structures in the image to generate the words 'Parliament Buildings'. For table visualization, special tokens have been removed.

Figure 6: Qualitative results from evaluation on WIKILANDMARKS. We observe that VT3 (w/o Vision) misinterprets the underlying landmarks. For example, It misinterprets 'Capitol Park Hill District' as a park and 'Conne Island' as an island. However, with the help of vision augmentation, VT3 correctly interprets the landmark types.

5.3 Human Evaluation

To better understand the effect of using visual information in VISTOT, we conduct a human evaluation of the results obtained by VT3 and VT3 (w/o vision). To this end, we randomly sample 200 generated sentences by each model on the WIKILANDMARKS test set and have them evaluated by three human evaluators. All these evaluators are in the age group of 20-25, years including two males and one female. They all have undergraduate degrees, are proficient in English, and are travel enthusiasts. They were all compensated appropriately for their contributions. The annotators were tasked to evaluate the generated sentences on the following metrics: (i) **Fluency**: measures coherence and grammatical correctness of a candidate sentence, and (ii) **Faithfulness**: measures whether all the facts stated in the candidate sentence can be inferred from the image or table. Fluency and Faithfulness are measured on a 3-point Likert scale (0, 1, or 2). The evaluators are presented a candidate sentence with the table as well as the accompanying image. To avoid any bias, the sentences are presented randomly. Detailed annotation guidelines are in Appendix A.

Model	Fluency	Faithfulness
VT3 (w/o vision)	1.83	1.66
VT3	1.84	1.78

Table 8: **Human evaluation results.** Three English-proficient graders evaluated fluency and faithfulness scores for 200 randomly-picked generated sentences on a 3-point scale (0, 1 or 2), and mean scores are reported.

We report the human evaluation results in Table 8. We observe that sentences generated by both models are highly fluent (1.83 and 1.84). Interestingly, VT3 (w/o vision) fails to perform comparably with VT3 on faithfulness (1.66 and 1.78).

Our analysis is that the vision augmentation helps generated text remain faithful, particularly when the information in the table is sparse or noisy, thus preventing hallucinations in text, e.g., an image of a district confirms the relevant related facts in the table even when the table and title might weakly suggest it is a park (Figure 6). We analyze this further with a categorical error analysis on the generated text sentences in Appendix F.

5.4 Qualitative results and Visualization

Figure 6 shows qualitative results generated by VT3 and VT3 (w/o vision). We observe that the full model can generate more accurate and expressive sentences, incorporating information from tables and images. Further, to learn what VT3 attends to in the image and table during generation, we provide a visualization of the attention heatmap over the image and table during the generation of a sentence in Figure 5.

6 Conclusion

We have highlighted the possible benefits of leveraging multimodal data for better sentence generation by introducing a novel task, viz. VISTOT and a large-scale accompanying dataset namely WIKILANDMARKS. We presented a competitive model – VT3 that is trained using three novel task-specific pertaining strategies, namely, image-table matching, masked value modeling, and image captioning. VT3 shows impressive performance gain over those models which only rely on tabular data for table-to-text generation. In general, inspired by our proposed novel task and dataset, we look forward to exciting future research on bridging multimodal cues for more precise NLG.

7 Limitations

We note a few limitations of this work: (i) The proposed dataset provides a nice testbed for VISTOT in the tourism domain. In this paper, we showed significantly improved accuracy using VISTOT using travel images. We believe these accuracy improvements should generalize to other domains with rich visual cues. For example, (a) In healthcare, automated generation of short reports given medical images and tabular data from medical tests can be modeled as a VISTOT task. (b) The other potential domains could be e-commerce, sports, and dining. Nevertheless, empirical validation needs to be carried out. We plan to therefore expand to more domains in the future. (ii) Richer context for table-to-text: The proposed approach (VT3) takes a table and a corresponding single image as input. However, multiple images may capture different aspects of natural language description in many scenarios. Similarly, other forms of multimodal context could be explored while extending the table-to-text framework further. (iii) Our dataset has a higher representation of landmarks belonging to countries such as England and the USA, which may lead to unintentional regional biases. (iv) Human evaluation could not follow a few best practices due to resource constraints, such as using a 7-scale Likert scale or evaluating a larger sample size. We acknowledge that such measures could have yielded more conclusive outcomes. (v) Our current work has focused on generating English sentences; in the future, we would like to extend this work to generate non-English language sentences. Especially the utility of visual cues for generating text for low-resource languages from tables and associated images seems an interesting research avenue. (vi) Given a table and an image, a human could generate a personalized summary s of information relevant to them. In this work, we modeled VISTOT as a problem of maximizing the probability of reproducing an existing sentence s', given the table and image. It is unclear how similar s and s' would be in real use cases. Again, due to resource constraints, we modeled the VISTOT task as a machine learning task of optimizing for s' rather than gathering s at a large scale.

8 Ethical Concerns

No personally identifiable data has been used for this work. WIKILANDMARKS data creation has been explained in detail in the main paper. Also, note that the images are all a part of Wikimedia. Therefore, most of these images are available under a Creative Commons Attribution license. All these images are publicly available on the web and may have different licenses.

9 Acknowledgements

We thank the anonymous reviewers and the meta-reviewer for their time and encouraging feedback. We thank Microsoft for supporting this work through the Microsoft Academic Partnership Grant (MAPG) 2021.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 6077–6086.

Gabor Angeli, Percy Liang, and Dan Klein. 2010. A simple domain-independent probabilistic approach to generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 502–512.

Junwei Bao, Duyu Tang, Nan Duan, Zhao Yan, Yuanhua Lv, Ming Zhou, and Tiejun Zhao. 2018. Table-to-text: Describing table region with natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020a. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Andrew Chisholm, Will Radford, and Ben Hachey. 2017. Learning to generate one-sentence biographies from wikidata. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 633–642.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from rdf data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does bert with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275. Association for Computational Linguistics.

Percy Liang, Michael I Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 91–99.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23.

Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, et al. 2021. Dart: Open-domain structured data record to text generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447.

Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. 2017. The e2e dataset: New challenges for end-to-end generation. In *18th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 201–206. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021a. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021b. Learning transferable visual models from natural language supervision. *Image*, 2:T2.

Ehud Reiter. 2007. An architecture for data-to-text systems. In *proceedings of the eleventh European workshop on natural language generation (ENLG 07)*, pages 97–104.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Somayajulu Sripada, Ehud Reiter, and Ian Davy. 2003. Sumtime-mousam: Configurable marine weather forecast generator. *Expert Update*, 6(3):4–10.

Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021. Plan-then-generate: Controlled data-to-text generation via planning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 895–909.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Yiyu Wang, Jungang Xu, and Yingfei Sun. 2022. End-to-end transformer based model for image captioning. In *AAAI*.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2575–2584.

Sam Joshua Wiseman, Stuart Merrill Shieber, and Alexander Sasha Matthew Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP2017)*. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Rong Ye, Wenxian Shi, Hao Zhou, Zhongyu Wei, and Lei Li. 2020. Variational template machine for data-to-text generation. In *ICLR*.

Zhipeng Zhang, Xinglin Hou, Kai Niu, Zhongzhen Huang, Tiezheng Ge, Yuning Jiang, Qi Wu, and Peng Wang. 2022. Attract me to buy: Advertisement copywriting generation with multimodal multi-structured information. *arXiv preprint arXiv:2205.03534*.

Appendix

A Human Evaluation Guidelines

The evaluators were provided with the following guidelines for annotation.
Goal: Evaluate the natural-language generated sentences on the following metrics.
Metrics:

- **Fluency**: How fluent and understandable is the generated sentence? Score on a 3-point scale: Not-Fluent, Mostly-Fluent, Fluent (or 0, 1 or 2 points)
- **Faithfulness**: Does the generated sentence have facts only inferred from the table and image? Score on a 3-point scale: Not-Faithful, Mostly-Faithful, Faithful (or 0, 1 or 2 points)

A.1 Example #1

Key	Value
name	Colwell Bay
location	Isle of Wight
interest	Geological
area	13.56 hectare
notifydate	1959
map	Natural England

Table 9: Table for Example #1

Figure 7: Image for Example #1 and #2

Generated Sentence: Colwell Bay is a bay on the west coast of the Isle of Wight.
Expected annotation: Fluency=2, Faithfulness=2

A.2 Example #2

Key	Value
name	Colwell Bay
location	Isle of Wight
interest	Geological
area	13.56 hectare
notifydate	1959
map	Natural England

Table 10: Table for Example #2

Generated Sentence: Colwell Bay is a 13.56 hectare biological Site of Special Scientific Interest on the Isle of Wight.
Expected annotation: Fluency=2, Faithfulness=1

A.3 Example #3

Key	Value
name	Taunton Castle
location	Taunton , Somerset
caption	Taunton Castle
map_type	Somerset
type	Norman
coordinates	51.0158 N, -3.1046 W
built	1129
builder	William Giffard
materials	Stone

Table 11: Table for Example #3

Figure 8: Image for Example 3

Generated Sentence: Taunton Castle is a Norman castle in the village of Taunton, Somerset, England.
Expected annotation: Fluency=2, Faithfulness=2

B Implementation Details for Reproduciblity

VT3 is implemented using the Hugging Face library (Wolf et al., 2020). The backbone of VT3 is the pre-trained $\text{BART}_{\text{base}}$ model, and the vision backbone is a frozen $\text{Swin}_{\text{large}}$ (Liu et al., 2022) model pre-trained on ImageNet-21K at a resolution of 384×384. Extracted visual embeddings for an image consist of 144 features having a dimension length of 1536. In total, VT3 has 339M parameters (including 197M belonging to Swin). We optimize training using the AdamW optimizer (Kingma and Ba, 2015) with a learning rate of 2e-5, a warmup for 2000 steps from 1e-7, and a batch size of 200 with a gradient accumulation value of 2. The model is trained for about 120K steps on four NVIDIA A100 GPUs, which takes $\sim$48 hours to complete the training process. For details, refer code[4].

[4] https://vl2g.github.io/projects/vistot

Location	Percentage
England	22.5
USA	10.0
Norway	7.0
India	5.7
Japan	3.3
Sweden	2.6
Italy	2.6
Australia	2.0
Germany	2.0
France	1.9
Philippines	1.9
China	1.7
Denmark	1.5
Canada	1.5
Others	33.7

Table 12: Country-wise distribution of landmarks in WIKILANDMARKS for the top 15 most frequent countries.

Location	Avg. BLEU
USA	31.1
England	32.4
Germany	34.5
Canada	28.1
France	30.1
Italy	28.8
Australia	36.8
Japan	34.4
India	20.4
Switzerland	42.7
Others	27.3

Table 13: Performance of VT3 across the top-10 frequent locations in the test set of WIKILANDMARKS.

C Examples from WIKILANDMARKS

We provide a selection of examples from WIKILANDMARKS in Figure 9.

D Geographical Distribution of Landmarks in WIKILANDMARKS

To understand where the landmarks in the WIKILANDMARKS dataset are geographically located and their distribution, we provide information regarding the most frequent landmark locations in Table 12.

E Analysis of the performance of VT3

We measure the performance of VT3 across the following dimensions of landmarks: (i) Location (in Table 13), (ii) Type (in Table 14), (iii) Architecture (in Table 15), and (iv) Material (in Table 16).

F Error Analysis of VT3 Generated Sentences

We perform detailed error analysis on 100 randomly selected test samples and compare VT3 and VT3 (w/o Vision). The error categories found from the results and their definitions are as follows – (i) **Incorrect Landmark Type**: model fails to interpret the landmark type correctly. (ii) **Fact undercoverage**: model fails to use all relevant facts from the table or image. (iii) **Fact hallucination**: generated sentences contain facts that are not grounded in table nor image. (iv) **Grammar errors**: gen-

Landmark Type	Avg. BLEU
Castle	28.6
Reservoir	34.6
Stratovolcano	20.6
Public	28.3
Protected	40.7
Art museum	39.7
Settlement	11.3
Urban park	21.8
Public park	23.5
Lake	31.4
Others	30.2

Table 14: Performance of VT3 across the top-10 frequent landmark types in the test set of WIKILANDMARKS.

Architecture	Avg. BLEU
Greek Revival	30.0
Classical Revival	38.6
Gothic Revival	42.2
Georgian	28.2
Federal	35.7
Italianate	31.1
Gothic	41.8
Late Victorian	45.9
Romanesque	36.3
Colonial Revival	33.5
Others	30.0

Table 15: Performance of VT3 across the top-10 frequent landmark architectures in the test set of WIKILANDMARKS.

Material	Avg. BLEU
Steel	27.9
Limestone	15.4
Stone	30.2
Concrete	32.8
Bronze	34.2
Reinforced concrete	22.8
Brick	28.4
Marble	18.2
Granite	33.6
Prestressed concrete	22.2
Others	30.2

Table 16: Performance of VT3 across the top-10 frequent landmark materials in the test set of WIKILANDMARKS.

Error Category	VT3	VT3 (w/o vision)
Incorrect Landmark Type	7	24
Fact undercoverage	18	26
Fact hallucination	6	10
Grammar Error	8	12

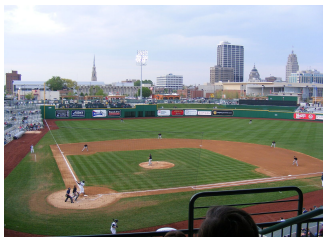
Table 17: Results of the error analysis on 100 randomly sampled generated test sentences from VT3 and VT3 (w/o vision).

Name	Niesen
Elevation	2632 m
Prominence	407 m
Isolation	2.3 km
Parent Peak	Albristhron
Location	Canton of Bern, Switzerland
Range	Bernese Alps
Coordinates	46.64611°N 7.6525°W
Easiest Route	Niesenbahn

The Niesen is a mountain peak of the Bernese Alps in the Canton of Bern, Switzerland.

Name	Amitabha Drukpa
Map	Location within Nepal
Map Size	250
Location Country	Nepal
Location	Kathmandu
Dedicated To	Amitabha

Amitabha Monastery is a Tibetan Buddhist Monastery in Nepal



Name	Parkview Field
Location	1301 Ewing St. Fort Wayne, IN 46802
Coordinates	41.07406°N -85.14286°W
Broke Ground	December 26, 2007
Opened	April 16, 2009
Owner	Hardball Capital
Surface	Kentucky Bluegrass
Cost	$30.6 million
Architect	Populous
Tenants	Fort Wayne TinCaps (2009-present)
Seating Capacity	6,516 (Fixed seats) 8,100 (Total)
Dimensions	Left Field – 336 ft, Center Field – 400 ft, Right Field - 318 ft

Parklow Field is a minor league baseball stadium located in the central business district of Fort Wayne, Indiana, US,

Figure 9: A selection of examples from WIKILANDMARKS. Please refer to Section 3.2 for more details.

erated sentence contain misplaced, incorrect, or redundant words, or incomplete sentences. We present the findings in Table 17.